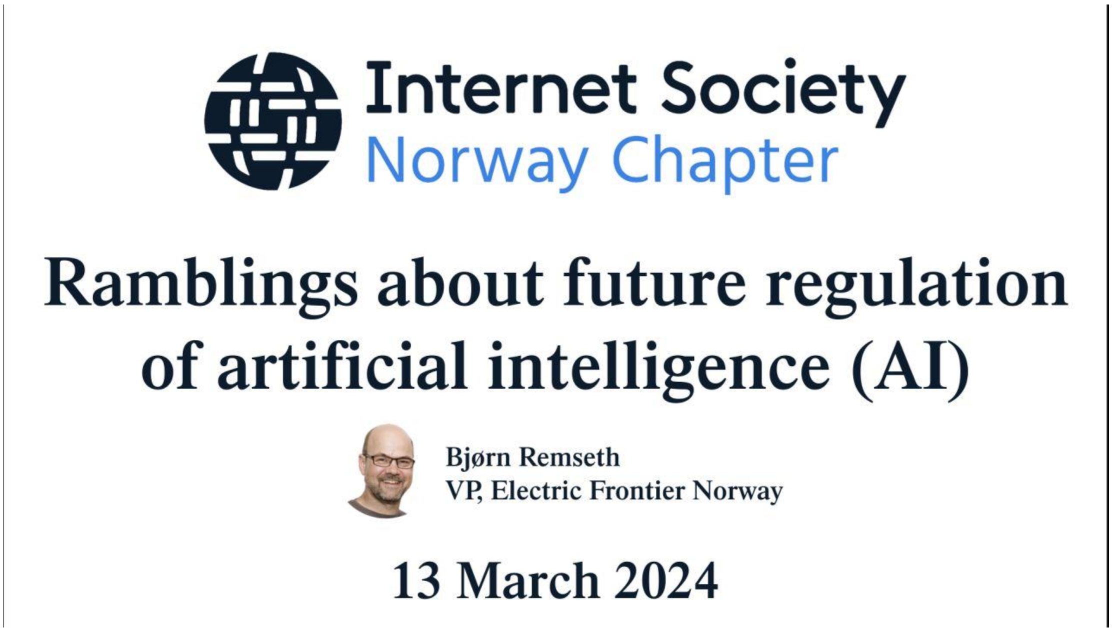

Internet Society
Norway Chapter

Ramblings about future regulation of artificial intelligence (AI)

Bjørn Remseth
VP, Electric Frontier Norway

13 March 2024

Ramblings about future regulation of artificial intelligence (AI) - Bjørn Remseth

ISOC Norway – 13 March 2024

Bjørn Remseth: Thank you so much. I'm the vice president of an organization in Norway called Electronic Frontier Norway. This is just my little disclaimer. What I'm about to present are ramblings, they're not the EFN policy.

I am doing this in my role as the vice president of EFN. Which I'm very proud of. And we have internal discussions. This reflects some of that internal discussion, but again, not policy. So just so that we're clear on this, don't hang me. And if I'd say anything wrong, I will be so happy if you correct me.

Here's what I will talk about today. I'm going to talk about regulation of AI. I'm going to talk about how to regulate it. I'm going to sequence it, like, what are we going to regulate? I'm going to look at what we call AI today, and how to regulate that, and then something about what we can reasonably expect to emerge in the relatively near future. Maybe also somewhat less reasonably. Turns out that those two things sort of merge together. And then... I'll just point out right now that in the far future, all bets are off. We don't really know how to regulate that, but I do have some comments about that too.

Bjørn Remseth: Again the highlights, or not the highlights, but the headings about what I'm going to talk about is today, it's the laws that we have today, or the laws that we know will come, and I'm going to take a European perspective on this. Those are the ones that will be used to regulate. It's the AI Act, the GDPR, well, the Digital Markets Act I didn't list that here, but I'll list it later. And yes, I'm going to talk about those because they're actually pretty good legislation, and they're good frameworks for thinking about something as complex as AI. They also have some interesting consequences in the near, somewhat uncertain future, because we will see that the assumptions that went into making that legislation, they don't really hold even today and certainly not tomorrow.

First I'm going to say a little bit about how did we get here? Just a bit on how we get to go to AI. So, what are the drivers for what we call AI today? Probably, I guess, the most stable driver is hardware. We are getting more bang for the buck, and we've been doing so for as far back as we can remember. There's a question about how far into the future we can take this. It's exponential growth. At some point in the future it needs to stop, because that's what all exponential growth does. However, it hasn't really stopped yet. It has slowed down, but it hasn't stopped. So I'm just going to assume that it's going to continue into the foreseeable future, which is at least six months, maybe three years ahead.

Intel indicated one nanometer nodes in 2027, we're at three nanometers basically today. That in itself is an incredible increase. That's not three, it's a factor of nine or something like that. It's a lot. That's happening.

Then there are algorithms. Things are happening in the algorithm space, too, I'm just going to very broadly outline. So, we have this classic machine learning thing, with the classification regression, and this graph here shows the increase over time in performance of algorithms, or actually algorithms combined with hardware, but it's the algorithms that run on the hardware.

As you can see, this thing started out pretty early and some tests were made that made it possible to compare algorithmic and human performance. Algorithms were pretty poor. Humans were the baseline up at zero, and then we see for these tests, when you have these types of things that when you can really do comparisons, machines have done pretty well. These are algorithmic improvements. They are on top of the hardware.

Then, there's something else, you can't just classify things and see which direction is this thing going, which word am I listening to?

You can actually generate stuff too. Here's just a little slide showing that. That is also becoming much, much, much better, and there are algorithmic improvements that mainly drive this. In addition to hardware, obviously, but you just couldn't do this 10 years ago, but you can do it today. You couldn't do it 10 years ago, even with all the hardware in the world, but you can't do it today with a reasonably sized laptop, actually.

So, this is new, this is a way for machines to learn how to do things by looking at examples, essentially. There are variations of this, but it's not the same as these others. It's another class of algorithms, where you get things to learn. In all of these areas, there are improvements happening. At the same time, they're feeding off each other. They're not independent, but they're not also the same. So, I'm just pointing out that this is happening and it's not stopping. None of them.

I will not say much about economics and finance, because I don't know that much about it. I found a nice graph. The interesting thing here is that, well, this is the total addressable AI market from some source. If you look at the scale on the left side, it's in trillions. So, okay,. they assume that we're now at a little bit less than 2, and in 2030 we might be at somewhere like 12.... 15 trillion dollars. I can't vouch for these things, but, if people believe them, then money will continue to pour into these things, and you will get whatever you need to drive these other trends at the speed that they can humanly and non-humanly go at.

So, that's happening. Again, probably not true, but the numbers are big and will get bigger. That's my main point.

So here's philosophy. I just want to point out that there are some things I don't understand, and I don't know that anyone else understands them either. I haven't seen anyone who's been able to give a sharp definition of either what intelligence or what consciousness is. We can cite examples, but you can't give a strong, sharp definition that says what's inside and outside.

Last year a bunch of researchers at Microsoft Research wrote this paper with this provocative title 'Sparks of Artificial Intelligence: Early Experiments with GPT 4".

What they claimed was that, well, this thing isn't intelligent, but if it were intelligence, it would be doing things like this. And it would look like that. And maybe this are things that could be extrapolated into that. We don't really know. That's my main point. We don't really know. I would really like to say that, that we knew, but we don't.

So that's the background. The thing is moving ahead and our basic knowledge what it is doing, at the very foundation, what is intelligence? What is consciousness? What is this going towards? We don't know what that is. We don't know it, when or if we're ever going to hit it. Maybe you will, maybe you won't. I don't know. Anyway, so back to regulation.

Well, before I get to regulation, this is one of the more nutty projections. This is a guy called Ray Kurzweil, who made these projections about how much compute thinking power is in the world. He made this assumption, he made it this many years ago, like 20 years ago or something, that in 2050 there would be more compute power in machines than in all of humankind.

We're not on track exactly, but we're not completely off track either. So, he may be wrong, but it may not be completely wrong. I certainly wouldn't bet everything against him. But, I don't know.

So, regulating today's AI. What is it? It's so many things, too much to fit on the slide, and it's going to get more because of these drivers. How on earth are you going to regulate that? I'm going to take a pretty close look at the EU AI Act, which I like a lot. It's not perfect, but it's certainly not bad. And sprinkle it with something about some of the other things, but mainly from today's perspective, right now.

This is a nice figure from the EU commission. It's a report sent to the commission. There's a link somewhere. They look at systems with various degrees of capabilities, and various risks while using them, and they try to address them independently, and I think they do a pretty decent job of that. We can criticize it, but that is in itself a good thing, right? Because it's so clearly defined that you can criticize it.

I'm just going to walk through this pyramid one step at a time, and talk a little bit about each one of them.

There's the briefing doc that I have stolen this from. It's a good briefing doc. You should look at it if you can.

They are assuming that, at the bottom, there's going to be a bunch of general purpose AI models. That is probably going to be true for a while now, probably, I mean, big models like GPT, LLaMA, these are general models that can be used in many, many respects. This is not wrong from today's perspective.

Bjørn Remseth: If you make these general purpose models, there are some things you must do. You must have documentation for them, and you must be able to make this available to people who depend on them, downstream providers of AI systems based on these general purpose models.

Also, you need to declare that these things that you are making will respect the EU law, including copyright, GDPR and all the rest of it. You must do that, otherwise you will not have a general purpose AI model that is fit for EU consumption.

Also, and this is interesting, they had this idea that if you have something that is really capable, like having 10 to the 25th FLOP capability, you must warn the EU, because you have a system that's so capable that, if general AI or something appears, it might be there and then it might run away, so you should know where it is or something like that.

Also, there will be cybersecurity issues with these things. That's just a fact of life these days. Any system that's sufficiently important will have cybersecurity issues, so you must be constantly vigilant, and you must do mitigation all the time, otherwise you're not in this business.

So, those are the systems that they assume everything will be based on. I will get back to why that's not really true, but that's further down the road. They classify them into minimal risk, something called transparency risk, high risk, and unacceptable risk systems. I'll define what they are, and tell a little bit what they need to do.

There are things like minimal risk, which is basically things like simple classifiers, like spam filters and things of that nature. You must comply with law, but there's nothing much about these things, if you have something that can be classified as minimal risk. I will point out that probably almost every interesting system will probably not be minimal risk, so this is probably a very small percentage of those systems that are, at least, being imagined.

Then they have something they call transparency risk, which is very interesting, because AIs can impersonate people, and, if you are in some kind of interaction with something. You shouldn't have to guess if it's a bot in the other end. That must be declared.

And also, if you make synthetic content, or at least lots of it, so that you have this huge corpus of something which looks like it's from real people, it must be clearly marked that it's not, so that you don't do silly things with it. And, by the way, synthetic data is going to be hugely important for the reason which I will get to in probably the next slide which is about high risk systems.

High risk systems. That's, as far as I can tell, almost everything interesting, because there's not that much that you need to do to be called a high risk system. So, there might be some law class working with specific sectors in the industry, for instance, medical systems, that can say, well, this system controls this thing, it's clearly high risk. If this insulin pump figures out that today we're going to change your dosage, and that's not to your benefit, that is high risk. That is kind of the old school high risk technology systems.

Then they have what I think is almost a new one, and that is any system that make profiles on natural persons will be considered high risk. This somewhat overlapped with GDPR, which has to do with information about people, but if you have a system that reads your biographies from somewhere, or biographies of multiple people, and then tries to make inferences about you, or a group of people, or how those people interact or whatever, then you are, by this definition, a high risk system, and you get a whole raft of requirements on you if you want to let that loose on people.

Which I think is good, but, oh boy, that's going to be challenging, because there are so many, and they have some procedures that you need to run through. I don't know what they are, I just read their word, but you can't make these systems, these high risk systems, and then just send them out, you need to, at the very least, register them.

And then, there are some provisions for making that easier and, if our friend Mika had been here, he could tell us a lot about that now, because he was part of the finalization of this legislation in the EU, but I don't know those details, unfortunately. But, there is a process that I know, and that is high risk. I'm just going to read this because it's."

anything that can cause significant harm", basically that's an AI system, anything will be prohibited.

So, if you try to manipulate people using subliminal techniques, out. If you try to exploit vulnerabilities, disability, you might have various types of disorders, if you try to target people based on that, that's out. If you try to use biometric to make inferences, for instance, you look like a rebel, you look like someone we should keep an extra eye on, that is, by default, not okay, it is prohibited.

Yeah, question?

[Participant]: Yeah, when you say prohibited, is this prohibited in the same way as in DDR, where the government's laws allow... let's say this is prohibited by default, but if the government decides that the police can use such techniques, it's still allowed?

Bjørn Remseth: There are exceptions for some of these.

So, for these high risk systems, for instance, if you use them to read emotions -- there's one down here -- if you try to infer emotions, that is illegal, but you could do it in a medical setting, so there are some exceptions there. It's not a blanket prohibition, but it's pretty much a blanket prohibition.

All of these things that are called -- what is it in GDPR? -- particularly sensitive information, like race, political opinions, trade unions, sex life, et cetera. All of those, if you do anything with them in an AI system, you are in prohibited area.

I wonder how that's going to play out for the dating apps, when they want to use this? Seriously, this is a law that they need to understand, and their system need to be conformant to it, if they're going to operate in the EU. And, they will find a way to operate in the EU. So, these are things that we will discover in the years to come.

Also real time biometrics, for instance, which basically is that you have cameras that looks at bunches of people and say, Hmm, that person, we want to sell to them, that is just blanket illegal. I would guess that it's still going to be legal in some law enforcement settings, but we need to see how that plays out.

Another thing is that there are systems like this already, and they are in production, and this AI Act says that they need to be phased out within six months after the Act goes into force, so that's going to be real interesting.

And, by the way, I fully approve. These are really obnoxious practices, and the fact that they are getting regulated, in my opinion, is good.

Steinar Grøtterød: There is a raised hand from this room.

Jayachander Surbiryala: Hi.

Actually these prohibited AI, whatever you have written here in this slide, these are examples provided by EU in the act, or you have written it as an example here? You have given biometric categorization, facial recognition databases. These are two aspects I'm curious about because, in this case, if you take the facial recognition databases, even EU has border control with facial recognition, which becomes illegal the moment they start to use AI, because this is going to contradict their own security in a way.

So, my question is here, is it they clearly specified this facial recognition as an example in the EU Act, or you have written it up? That's what my question is at the moment.

Bjørn Remseth: Okay. So, I paraphrase this from a briefing document for the Commission, the one I gave a link to.

Jayachander Surbiryala: In that case, okay, maybe that they have mentioned it, and again, if you go back to two slides where you have talked about health, I think it was in high risk, you have talked about medical services, but medical service also leads to the biometrics, then they should not use the AI in medical areas as well. I think it is a bit contradicting, in some ways.

Bjørn Remseth: Tom Fredrik, here, has an answer.

Tom Fredrik Blenning: Interestingly enough, this was on a meeting last week that you were not in, Bjørn. When it comes to border services, and when it comes to law enforcement, there are carveouts for them. There are still talks about exactly which systems will be included, and we hope to see that these border security systems will be included, but, let's just say that the hope that this is going to happen is quite bleak at the moment.

Jayachander Surbiryala: Okay. Yeah. Thank you.

[Participant]: Yeah, it's one of the same questions I had because to me it's obvious that government can and will [inaudible] what they need to do for lawful use in their own right.

It sounds to me like it's in GDPR, where the government can do stuff that they don't allow the private sector to do, if it's not moralistic.

Bjørn Remseth: Yeah that could be it, yeah. But, at least you will not have the private sector doing it for their own purposes.

[Participant]: Basically, wasn't it a company called Clearview?

Bjørn Remseth: Yes.

[Participant]: They did make this on their own initiative and to shop it around, which would be clearly illegal....

Steinar Grøtterød: Nevermind, just proceed.

Bjørn Remseth: Okay, so he was just talking about the company Clearview, who did develop a database for doing facial recognition by just looking at a bunch of data on the internet. If that had been a European company, that would have been prohibited from this thing by this legislation.

So, there are some practices, there's a bunch, if you don't comply, you will get fined up to 30 million, or 6% total worldwide annual turnover. I don't think 30 million is that for Meta, so this is disappointing.

They also do encourage nations to develop sandboxes where risky systems can be tested out with very close scrutiny from regulatory regulators, and the reason for this is that it's really difficult to test out systems without having them in something very similar to real production, which is, again, why I personally believe that synthetic data is going to be hugely important, because if you want to develop something that is potentially risky, setting up a sandbox, or something which is within the regulatory constraints, just to do your initial development, it's just going to be prohibitively expensive.

So, we're probably going to see a lot of development happening on synthetic data, and then, at some point, it will need to go through these steps to get approved. That's just a guess on my part, but it's an educated guess. At least I hope it is.

That actually concludes the AI Act, what I'm going to say, and, by the way, that is one that has the most slides at all, so there's going to be a lot fewer slides per thing from now on.

One of the things that is of importance is the European Digital Markets Act that's coming online as we speak. They try really hard to bring in a lot of online platforms, and their gatekeepers, platforms like browser, search, video, social networks, et cetera, and the gatekeepers, which are the big companies controlling them. They are going to enforce stricter rules for consent.

I don't know how that's going to play out, but it's going to be more than just consent for this particular website. if each data is going to be shared, my understanding is that you will need extra consent for that. Also, more interoperability in the sense that you can say, well, I don't want to use this particular search engine, I want to choose whatever I want. You will be, by law, have that right. And also, there's going to be a few other things like, you have the right to uninstall whatever you want.

How is this going to play out on the AI side? That is unknown because, in this list, AI technologies are not listed, because, well, we have a separate legislation for that, and this law, well, we just need to push it through, so we can't mix it up with the AI, I guess. But, for sure, that's going to happen because you're going to have models that do things These things will be of importance for you, and if you have something, say, from Meta that is good, and you want to use it in a Microsoft platform, why should Microsoft stop you from doing that, or vice versa?

In some way, it's the same thing as with search engines and browsers, but this is going to take a while to play out, because it's not the same, it's just sort of the same. In principle, it is the same, in practice, it's very different. And again, the law wasn't made for this particular case, so it's going to be played out in the market and the courts and the lawyers, but it is a relevant piece of legislation.

Here's... I introduced a new symbol at top, it's a cherry, it's a cherry that's being picked because there are so many things to talk about here that I just need to cherry pick things, I have no hope of covering the whole thing, so I'm now cherry picking one thing, and that is that, European Convention of Human Rights, that is also going to be relevant here.

I'm going to spell something out. Article 10, my favorite, that gives you the right to information. It gives you freedom of expression. It also gives you freedom to receive information. Both. Very important. But, there's a question. A concrete example, if you make a language model, for instance, that you can ask questions, who has the right to censor that language model and to what extent, and when does that start to get into trouble with the Article 10?

For instance, I have a test I usually run, just for fun, I never intend to do anything with it, and that is I ask these language models how to synthesize various chemicals that are known to be dangerous, and it's interesting to see how they react. Some of them will give me a long lecture on the fact that I shouldn't use drugs, even if I'm just interested in certain parts of organic chemistry. Others will actually give me an outline of recipe. So, there's variation, and , these are from very serious actors that do this very differently.

Just today. I think I heard that Google were not going to let their latest language model answer certain questions about elections. If these things start to insert themselves into the streams of things that informs us, which they are doing now, at what point are they so important that these safety regulation, safety mechanisms, as they call them, become so prohibitive that they come in conflict with the human rights? I would guess that that would happen at some point, hasn't happened yet, but, again, this is legislation that I think will touch on AI. It's not going to happen today, but maybe this year, maybe next year, it's probably going to happen not very far into the future.

Okay, so that was today, and now I'm going to take the future. The future is going to be quite quick compared to this, but that's where we are all going to live, so it's important to spend a little bit of time on it right now. Again, I'm cherry picking. I'm going to cherry pick robots because I like robots.

We're going to look at robots in human space, human space being defined as this, where we humans are, the space we occupy, the things we do. I'm going to argue that robots are going to be more and more present in that space. As a consequence of that, they're going to get much more information from that space. They're going to see how we put things, where you put things, how we interact. They're going to benefit from being able to model our interactions so they can fit themselves into that, both by avoiding hitting us and also being helpful for us.

But, oh boy, are they going to make profiles of us?

Yeah.

Let me see, did you mention that earlier? Yeah, we did. That is by definition a... what was it? A high risk system, yeah. So, any robot operating with an AI in your house is going to be high risk system by definition. Not going to avoid it. Probably won't be very useful, maybe, but yeah.

Why do I think this is going to be true? I'm going to just outline the argument.

20 years ago, I saw a very nice video by Andrew Ng, a very famous AI researcher. He showed two video clips, one of his helicopter that he had trained using an older type of machine learning, where he did an inverted funnel, very, very tricky operation with a radio controlled helicopter. But, that was something that he could do at that point. Then he showed another video, which I thought was even more impressive, where there was a robot that was rolling around in the flat, and picking up stuff and putting it in the right place. Then he said, well, this thing is actually guided by a graduate student.

His point was that, at that point, they knew very well how to do control tasks as in, so called, certain type of control theory, but they did not know how to do navigation in human space and planning in human space. That is no longer true. That is changing really fast now.

Here's an example, I don't know if you can see it, but it's a couple of robot's arm that's making an omelet, I think. It's broken the eggs and it made the omelette. There are examples of robots that can do this. They can learn it by looking at people doing it. These things are happening.

Humanoid robots. This is very silly one, but it's a tool helper. There are many companies that are making products there. This trend actually is. on the cusp of happening now. The timing is right. There are many companies betting on that, so it could happen.

Here's another thing which is not robotics, but I'm going to mention it with robotics because it is touching on the same things, it's the AR / VR thing. If you put a set of AR glasses on your head, augmented reality, they have cameras going in many directions. They make a map of your surroundings, and they place you in that surroundings. As you navigate, they have a model of your surroundings that they keep you in, and your views of it and the views of everything else that you can see, and some of the things you cannot see, because it has a wider range of view than you have. You can think of that as technology that guides the robot, but you are the control system for that robot.

The thing is that, if it has any kind of traction in the market, it'll drive up the number of units produced and down the cost per unit. That means that this trend will accelerate. It means that you will have very capable systems that are traveling around in your house, and getting a lot of information about you and everything in there. You actually don't need robots for that to happen, that's going to probably happen anyway. If AR / VR happens, these two things will feed on each other.

So, what are we going to do about this? I actually don't know. We will just need to see if we can map the current legislation onto this, because we will not have time to make new legislation, it's going to happen so fast.

I'm going to give you another example of something which is going to challenge this current legalization, and that is small models. You've all heard about large language models, but some of you may also have heard about small models. One of the interesting things that's happened the last not so many years, since this is really recent, is that a lot of research has very successfully gone into making models that are very capable, that are very small.

You use a larger model and then you, in a sense, compress it, or you use it to train a smaller model to do very well on the particular more specialized domain. You can get models that are actually quite small in physical size that can run relatively fast on consumer hardware today. That's not going to be worse in the future, it's going to be better.

That means that the assumption that you're going to have this basis of big models that are well regulated, that's not necessarily going to be true, because you're going to have very, very capable models running around in things of this size. I'm holding up my phone now. Just to emphasize that point in a small way, when Samsung announced the latest Galaxy phone, they made a point of it having one of these small models, that Google has made, in there. I don't know how useful it is, but it is a thing. It's not going to be less of a thing in the future, I think. These are relatively small. They ride on the hardware and algorithmic improvement trends that I outlined, and they can run a model starter inside robots and personal devices.

It's not going to be a few large owners. It's going to be really distributed ownership of it. How on earth are you going to regulate this?

Bjørn Remseth: Again, I don't really know. I'm looking into my crystal ball, and I'm seeing muddy things, but it's unclear. Here are some ideas. Maybe it's possible to define safe-self basis, so that maybe, if you are a robot producer, you have this model, it has been certified and it does only this very limited set of things. Maybe that's something you can approve. Maybe you can make it a tamper proof. That's going to be difficult if you also want a learning system, and these systems are, by definition, very useful if you make them learning systems.

It may be that it's primarily transparency, that is the way you want to go with these things, that you have a large set of properties about them that are known, so you can have... I don't know even which word to use, intelligent, sensible, reasonable set of assumptions about what the thing does, and so you can take responsibility for it, if that is possible?

So, if you get widespread ownership, who's responsible for these systems actions? If these systems do things, who is responsible? Is it you who own the system? Because if it is, then you are really in a bad situation, if these things start doing bad things on their own, and they can be autonomous, right?

So if not, and if you are responsible, how are you going to take on that responsibility? One classical way of doing this is to buy insurance, that's what we do with cars. Cars are really dangerous and really risky, but you can still go on the road because we have this contract known as insurance, and, as long as they have that insurance, we have a limited liability,

But, if you're going to have an insurance mechanism for AI, how on earth is that going to be formulated? That is a big opportunity for someone, because if AI happens in a distributed way with multiple people responsible for their actions, they will need insurance, and I have no idea about how that is going to happen.

I'm getting close to the end now. I do have a nice crystal ball as you can see. Basically, I just hope that the framework that has been produced can evolve and follow these things. It needs to do it, otherwise it will be uncontrolled. What if the big one drops, and what is the big one?

The big one is Artificial General Intelligence. That's basically a thing that is as smart as human in most domains, and it has agency and free will. What on earth are we going to do about that?

I'm just going to read this because I just thought it out not so long ago. Basically, we have no clue, it's fundamentally unknown. What is sure is that, if these things appear, we will need to co-exist with them, and they will in many ways be superior to us. If we want to regulate them, and we probably do, then it will need to be a collaborative effort with the AGIs.

How on earth are we going to make preparations so that we and the AGIs can collaborate? There are these people talking about how to do alignment, and so on. I don't know.

I personally have more trust in the EU, if the EU is able to concretely make regulation that makes it possible for AI and humans to evolve and co-exist. To my mind, that's probably a better guess than some philosopher sitting somewhere and figuring out something that may or may not work in the real world. E-regulation must work in the real world, otherwise it's useless.

That is actually my hope. I really hope that that type of regulation will be so good that actually the machines will think it's a good thing to continue in that direction.

So, we really must take it very seriously.

That was actually everything I had to say. Thank you for your attention. Here's the QR code that goes to the whole presentation. And if there are any questions, I will be happy to do my best to answer them.

Steinar Grøtterød: First of all, thank you, Bjørn. This was another view of things that I tried to understand and it's always good to have another view on it. So, my understanding what the EU is doing now is really trying to regulate this. , Really trying to regulate this, and we have spokesmen and politicians in Norway also said we need to regulate this.

My question is, is the ideal with the regulating is to define what sort of level of risk a certain program, a certain feature, is taking, and if so, how can we test this out? Is it by looking at the purpose of how AI is to be used, or is it the sandbox, or is it the code, or is it something else?

Bjørn Remseth: All of this actually needs to be discovered. We don't know. Basic principles, first try to avoid harm. Obviously, harmful things should be avoided, and, if they can't be avoided, they need to be tested and confined. How to do that? Not really known. At the same time, if we can't do experiments to see where we can go, then everything of the good stuff, there's evidently many good things that can happen here, we won't get to it, because we have blocked ourselves, because we reasonably see potential harms in every direction. So, this is a really tricky one.

I think we need more than sandboxes, actually. We need experimental areas where we can do almost completely wild experiments without harming anyone, and that probably means that we can't use much real data at all, and when we do, in these very experimental settings, it must be very, very tightly confined.

It will change the way we develop systems, I'm pretty confident.

Steinar Grøtterød: I do agree. My thinking is that we have heard that AI is the new industrial revolution, and that's kind of maybe a buzzword of popular things, but there is a challenge within EU on making laws for regulation about this. That's time consuming, and the speed of this technology is definitely ahead of the law making process here within EU.

Bjørn Remseth: Not only that, this is very much a case of practice being ahead of theory. That is true clearly for regulation, but also for technology. I mean, who can tell me in detail how the transformer model works? Not that many. It works, and we can say things about how it works, but we're going to spend many, many, many years doing theory on that one yet. We don't know everything about them yet, but we do have the phenomena, so there's this tension here.

[Participant]: My question is, regulation is fine, it's good to have, but isn't it a bit naive that we're going to be able to stop things from happening just by focusing on regulation? Let's take a parallel to viruses and malware and spam. That's been going on since the beginning of communication online, and we've regulated it for all these years, but has the problem disappeared, or has it grown? With these models being online on the Internet, with all the bad actors that can use them, isn't there something else that we also should be focusing on?

Bjørn Remseth: I obviously don't have any clear answers for that, but I do have one somewhat unclear one, and that is that enforcement needs to be done. If these transgressions are discovered, they must be reacted to, otherwise the system as a whole will not learn to avoid them. That's the best answer I have, but clearly it's not sufficient. I completely agree with that. It must be more detailed than that.

About general artificial intelligence. When that happens.

Tomas: We have to consider them, they will be like persons, they will have some kind of rights, maybe like children. I'm pointing towards the paradigm of parenting. If we are good parents, then maybe they will respect us, but if we are bad parents, dysfunctional parents, then maybe they'll hate us or ignore us, or maybe even try to kill us.

You are a parent, so could you say, what do you think about the parenting perspective concerning artificial intelligence?

Bjørn Remseth: I'm not so sure that's a good perspective. The reason is that you have a pretty intimate and very controlling relationship with your child, and with emerging AI there's absolutely no guarantee that that's going to happen. More likely than not, it won't happen. That's a whole different talk about, if we ever get to AGI, how is that going to happen? How is that going to emerge? Is it going to be kept secret? Is it going to happen in multiple places at the same time? Is it going to be state actors, bad actors that do it first and then control it, and so on and so forth? I actually don't think parenting is a useful paradigm for this particular case. I understand it, but I personally don't find it useful.

[Participant]: Can you elaborate on the exciting point about AI alignment, and how that compares, or is alternative, if that's what it is, to other ways of regulating the behavior of the AI?

Bjørn Remseth: If I understand correctly, the idea is that we'll program into the AI's patterns of behavior, so that they will not go off on their own. Their interest will be tied to humans. Now, to do that, you will need to formulate relatively clear criteria for what is good for humans. Good luck with that, is my first comment. But, again, it's not really an option, if you're going to have these things, because they are going to be very powerful, and if we don't know how to reign them in...

[Participant]: Almost sounds like the 21st century version of the Asimov robot laws, in a way, and some kind of vision that you can hard code these values that we'll never go beyond, and that's the way to regulate.

Bjørn Remseth: If that was indeed possible, that would be fine. As I said initially, there are some philosophical questions here that are just not answered, at least not as far as I can tell, and this is one of them.

I don't have a good answer, that's my best answer. Tom Fredrik, you probably need to get here.

Tom Fredrik Blenning: One of the challenges with formulating these principles, as you're saying, but a very common example is actually if you formulate that all AIs should be good to the planet. One of the corollaries of that would be to eradicate humanity because we destroyed the planet. It's a good principle, but it's very difficult to give it in a way that doesn't have unforeseen consequences.